Mapping Artificial Intelligence Applications in Requirements Engineering: A CiteSpace-Based Bibliometric Review

Guoxin Lin[1,2,3], Ronghai Wang[1,2,3,*], Hongwei Wang[1,2,3], Fen Cai[1,2,3]
[1]School of Mathematics and Computer Science, Quanzhou Normal University, Quanzhou, China
[2]Fujian Provincial Key Laboratory of Data Intensive Computing, Quanzhou, China
[3]Key Laboratory of Intelligent Computing and Information Processing, Quanzhou, China
**Corresponding Author*

Abstract: Artificial intelligence (AI) and requirements engineering (RE) intersect in two related research streams. One applies AI methods to RE tasks, whereas the other adapts RE practices to the particular demands of AI-enabled systems. This study maps the development of both streams and examines how their evidence bases connect. A search of the Web of Science Core Collection returned 1,021 articles published between 1 January 2000 and 31 July 2026. Screening of the bibliographic fields excluded 897 off-topic records, one retracted article, and two articles with an Expression of Concern. The remaining 121 records were analyzed with CiteSpace 6.4.R2. The analyses covered keyword co-occurrence and clustering, thematic timelines, citation bursts, reference co-citation, and country collaboration; representative records were also checked manually. The keyword network contained 240 nodes and 695 links. Seven main themes were identified: AI-assisted requirements analysis, RE for AI systems and quality assurance, LLM-assisted automation, systems design, elicitation and inconsistency detection, human-centered explainability, and AI-based modeling. The reference co-citation network contained 911 nodes and 952 links, but only 16% of its nodes belonged to the largest connected component. Under the primary setting, no keyword maintained a burst for two years. A separate one-year sensitivity analysis identified a burst for large language models from 2025 to 2026. Overall, the mapped literature has recently moved beyond task-specific NLP and ML toward generative and agentic workflows. Industrial validation and traceability, however, remain limited, and lifecycle governance and human oversight have not yet been adequately resolved.

Keywords: Artificial Intelligence; Requirements Engineering; AI4RE; RE4AI; Citespace; Bibliometric Analysis; Large Language Models

1. Introduction

Requirements engineering (RE) defines what a system should do, which qualities it should satisfy, and how stakeholder needs and constraints should be negotiated, specified, validated, and managed. Much of this work relies on natural-language artifacts created by people with different vocabularies, priorities, and levels of domain knowledge. Ambiguity, inconsistency, missing information, duplication, and changing expectations are therefore routine problems. These characteristics have made RE a long-standing application area for natural language processing (NLP), machine learning (ML), and other forms of artificial intelligence (AI).

AI support for RE has progressed from rule-based language analysis and supervised classification to transformer models, zero-shot learning, and large language models (LLMs). A systematic mapping of NLP for RE identified applications in defect detection, classification, traceability, modeling, and related tasks [1]. Recent studies assess user stories, generate specifications, decompose functions, and link requirements with design or implementation artifacts [2,3]. A review of generative AI for RE also reports activity across several stages of the RE lifecycle [4]. This article refers to such uses as artificial intelligence for requirements engineering (AI4RE).

The relationship also runs in the opposite direction. AI-enabled systems depend on data, probabilistic models, and changing operating conditions; their behavior is often difficult to specify using conventional software

requirements alone. RE for these systems must consider model performance, data quality, safety, fairness, accountability, explainability, human control, and post-deployment change. Existing mappings suggest that RE methods and tools are not yet fully adapted to these concerns [5], while broader software-engineering research points to lifecycle and operational difficulties beyond model construction [6]. Human-centered frameworks seek to identify information and explanation requirements before an AI-enabled design is implemented [7]. This direction is termed requirements engineering for artificial intelligence systems (RE4AI).

The two streams are analytically distinct but increasingly overlap. An LLM may, for example, help specify an AI-enabled system; the same study can then involve automation, traceability, explainability, and governance. Existing reviews synthesize particular techniques or RE activities, but they do not by themselves show how AI4RE and RE4AI are positioned within a shared bibliometric structure. Science mapping can reveal those structural relationships, provided that statistical clusters are checked against the papers from which they were derived.

Two methodological issues require particular care. A broad bibliographic query retrieves many records in which "requirements," "engineering," or "AI" has only an incidental or domain-generic meaning. Visualization should therefore follow, rather than precede, transparent corpus screening. In addition, a statistically distinctive cluster label is not an expert summary of every record in that cluster. Representative-paper reading is needed, especially when a cluster contains few associated records or the wider network is fragmented.

Against this background, the review addresses four questions:

1. RQ1: How has publication output on artificial intelligence and requirements engineering evolved from 2000 to July 2026?
2. RQ2: What are the principal knowledge clusters and application themes in this research domain?
3. RQ3: How have the major themes evolved over time, and which topics have shown recent keyword bursts?
4. RQ4: How are AI4RE and RE4AI distributed across requirements-engineering activities, and what research gaps remain?

The analysis combines a screened Web of Science corpus, CiteSpace networks, representative-paper checks, and an evidence-maturity matrix. Its purpose is to distinguish observable bibliometric structure from claims about technical or industrial readiness, while showing where AI4RE and RE4AI meet.

2. Materials and Methods

2.1 Data Source and Search Strategy

The bibliographic data were downloaded from the Web of Science Core Collection. We searched the topic field with TS= ((Requirements Engineering AND (Artificial Intelligence OR AI))), restricted the document type to articles, and set the coverage period from 1 January 2000 to 31 July 2026. An inclusive query was used because the corpus needed to represent two forms of research: AI4RE studies, in which AI, ML, NLP, or LLMs assist RE activities, and RE4AI studies, in which RE is used to define data, model, safety, fairness, explainability, and other requirements for AI-enabled systems. The search produced 1,021 records. They were exported as three plain-text files containing complete bibliographic fields and cited references.

2.2 Data Cleaning and Corpus Screening

The records were screened in two stages. Web of Science accession numbers and normalized titles were first checked for duplicates; none were found. Rules based on RE activities and AI technologies were then applied to titles, abstracts, and author keywords. Records flagged as likely false positives, together with non-standard borderline cases, were reviewed individually against their titles and abstracts.

Records were retained when they addressed either AI4RE or RE4AI. They were excluded when "requirements" referred only to general constraints such as energy, nutrition, materials, equipment, or occupational competencies; when AI appeared only as background or future work; or when no substantive AI–RE relationship could be established. One retracted article, two articles carrying an Expression of Concern, and 897 clearly off-topic articles were removed. The final corpus comprised 85 core studies and 36 conservatively retained borderline studies, for a total of 121 records (Figure 1). This was bibliometric corpus screening based on bibliographic fields, not dual-reviewer full-text screening.

2.3 CiteSpace Configuration and Analytical Procedure

The 121 records were analyzed in CiteSpace 6.4.R2 (64-bit Advanced). Time slicing covered 2000–2026 with one year per slice and July as the final month of 2026. After empty intervals were removed, CiteSpace displayed 2001–2026 as the non-empty range. Terms were taken from titles, abstracts, and Author Keywords (DE); Keywords Plus (ID) was excluded. An alias list consolidated selected abbreviation, spelling, and synonym variants, including AI with artificial intelligence, LLM/LLMs with large language models, and requirement engineering with requirements engineering.

One core node type was used in each formal run. The primary selection criterion was the g-index with k = 25; link strength was cosine and link scope was within slices. The keyword network was pruned with Pathfinder, and initial cluster labels were extracted with the log-likelihood ratio method. The largest clusters were then renamed after checking two or three representative titles and abstracts; when only two associated records were available, both were reviewed.

Keyword burst detection used a minimum duration of two years. Because early annual counts were sparse, a one-year run was retained as a sensitivity check and reported separately. The cited-reference network also used k = 25 and Pathfinder pruning. Runs with k = 15 and k = 50 were used to check the stability of maximum-connected-component coverage. An unpruned country/region network with k = 25 provided supplementary collaboration context. Table 1 records the final analytical settings.

Table 1. Data Source, Screening, and CiteSpace Settings

Component	Final specification
Source and query	Web of Science Core Collection; TS= ((Requirements Engineering AND (Artificial Intelligence OR AI))); document type: Article.
Coverage and scope	1 January 2000-31 July 2026; both AI4RE and RE4AI were eligible. The final corpus contained 121 records.
Screening	From 1,021 records, 897 off-topic articles, one retracted article, and two articles carrying an Expression of Concern were excluded. No duplicate accession numbers were found.
Screening method	Rules applied to titles, abstracts, and author keywords, followed by targeted record-level checks; bibliometric corpus screening rather than dual-reviewer full-text screening.
Software and slicing	CiteSpace 6.4.R2 (64-bit Advanced); 2000-2026; one year per slice; final month July 2026.
Terms and normalization	Titles, abstracts, and Author Keywords (DE); Keywords Plus excluded; selected abbreviations, spellings, and synonyms merged with an alias list.
Selection and links	One node type per run; g-index k = 25; cosine link strength; within-slice links. Keyword and cited-reference networks used Pathfinder pruning.
Labels, bursts, and checks	LLR labels were checked against representative records and manually calibrated. Burst duration was two years, with a one-year sensitivity run. Reference-network checks used k = 15 and k = 50.

Note. The unpruned country/region network was used only as supplementary context.

2.4 Interpretation and Reliability Checks

For each primary network, the numbers of nodes and links, density, and largest connected component (LCC) were recorded. Modularity Q described separation among clusters, and weighted mean silhouette described within-cluster consistency. These statistics were used as diagnostics of a particular network solution, not as evidence that the themes were objectively true or exhaustive [8,9]. Cluster size was not treated as research quality, node frequency was not treated as a quality score, and burst strength was not interpreted as a forecast.

Seven keyword clusters and four reference co-citation clusters were checked against representative corpus records or cited works. Co-citation labels were assessed against both the associated citing records and the most frequently cited references. Interpretations were qualified where only two supporting records were available or where network fragmentation restricted coverage. Run parameters, indicators, output filenames, and sensitivity decisions were retained in the project log.

3. Results

3.1 Corpus Screening and Publication Trend

Screening reduced the 1,021 retrieved articles to

121 records (Figure 1). The exclusions comprised 897 articles without a substantive AI–RE focus, one retracted article, and two articles carrying an Expression of Concern. Of the retained records, 85 were classed as core studies and 36 as defensible borderline studies.

Publication output was sparse during the first two decades of the search period (Figure 2). The non-empty annual counts were one article in each of 2001, 2012, 2014, and 2017; two in each of 2015, 2018, 2019, and 2021; and five in 2020. Output rose to 12 articles in 2022, 16 in 2023, 17 in 2024, and 21 in 2025. Those four complete years contributed 66 articles, or 54.5% of the corpus. A further 38 records were indexed between January and 31 July 2026. This partial-year value was displayed separately and was not used to calculate a full-year growth rate. Overall, 111 of the 121 records (91.7%) appeared from 2020 onward.

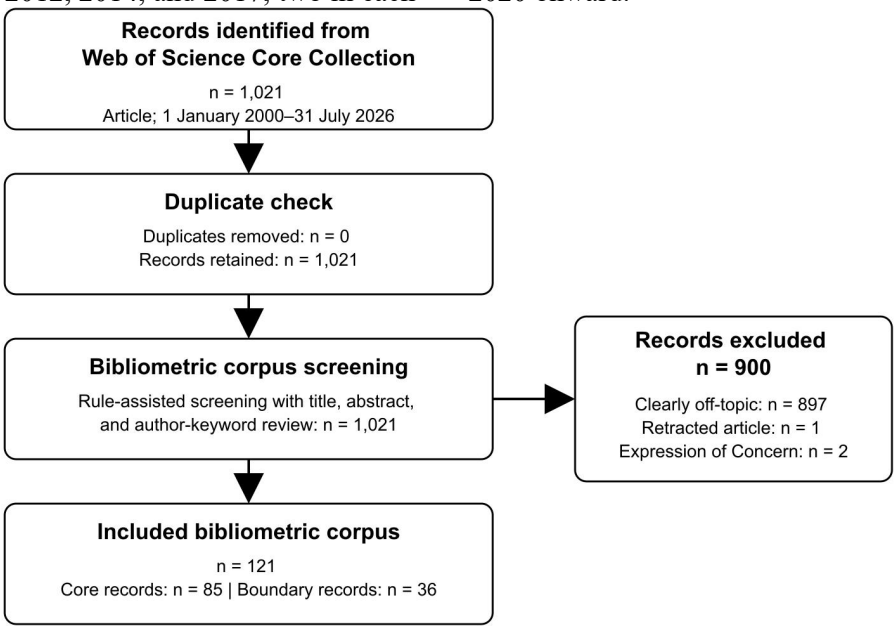

Figure 1. Identification and Screening of the Bibliometric Corpus

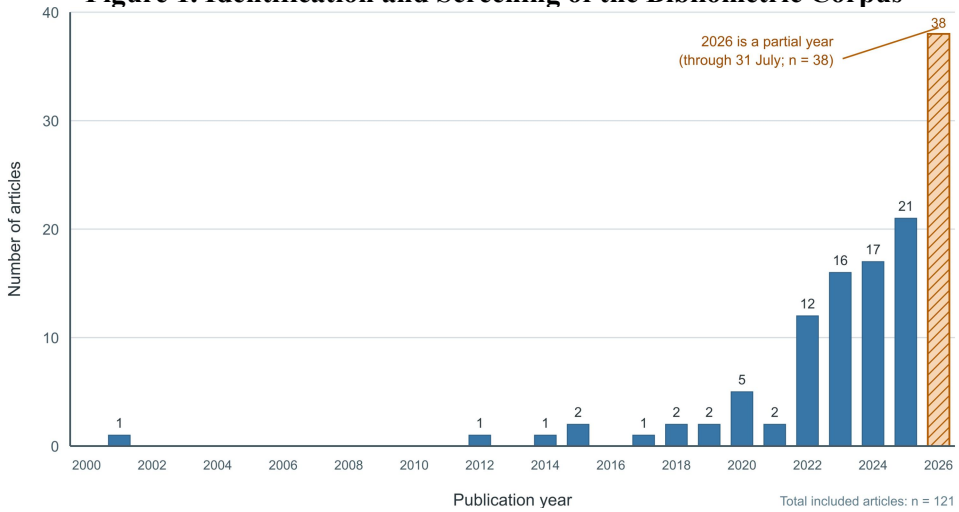

Figure 2. Annual Publication Output of the Included Corpus, 2000-2026

3.2 Keyword Co-Occurrence Structure and Thematic Clusters

The normalized keyword network contained 240 nodes and 695 links (density = 0.0242); its LCC contained 165 nodes (68%). The Pathfinder-pruned solution had a modularity Q of 0.6689 and a weighted mean silhouette of 0.8695. The most frequent normalized terms were requirements engineering (47 occurrences), artificial intelligence (35), large language models (24), machine learning (21), and explainable artificial intelligence (6). Their positions in the network core show the coexistence of discipline-level terms, established ML terminology, and recent LLM work (Figure 3).

The seven retained clusters together contained

165 nodes, exactly the size of the LCC. The other 75 nodes fell outside this principal clustered component. CiteSpace associated 51 corpus records with the seven cluster-information folders: 16, 15, 10, 3, 3, 2, and 2 records for clusters #0, #1, #2, #3, #4, #5, and #8, respectively. Representative-record checking covered three records for each of clusters #0–#4 and both available records for clusters #5 and #8. The labels therefore summarize the connected thematic core rather than all keyword nodes.

Cluster #0, AI-assisted requirements analysis, was the largest (size = 41; silhouette = 0.849; mean year = 2020). It combined language-model-based classification with explainability and other requirements of AI-enabled systems. Zero-shot requirements classification was one identifiable AI4RE component [10], while other records addressed RE4AI concerns; the calibrated label therefore preserves the cluster's bridging role.

Cluster #1, RE4AI and quality assurance (size = 38; silhouette = 0.773; mean year = 2022), had the lowest silhouette among the seven clusters. It included a mapping of RE4AI practices and challenges [5] as well as AI-assisted completeness and quality checking. Cluster #2, LLM-assisted RE automation (size = 34; silhouette = 0.866; mean year = 2024), was the most recent. Its records covered user-story assessment, specification generation, and requirements-to-development pipelines; comparative LLM-based user-story evaluation illustrates the quality-assessment component [2].

Cluster #3, AI-assisted systems design (size = 18; silhouette = 0.965; mean year = 2021), connected earlier work on ethical requirements for autonomous artifacts with agentic LLM support for requirements extraction, functional decomposition, and conceptual design [3].

Cluster #4, AI-based elicitation and inconsistency detection (size = 14; silhouette = 0.957; mean year = 2020), linked requirements acquisition from online forums with formal-logic and LLM-based contradiction detection [11, 12].

Cluster #5, human-centered design and explainability (size = 13; silhouette = 0.960; mean year = 2020), covered design-level classification and information requirements for explainable systems. Cluster #8, AI-based modeling and similarity detection (size = 7; silhouette = 0.940; mean year = 2022), covered AI-assisted domain modeling and semantic similarity among requirements [13]. Each of these two clusters had only two associated corpus records, so their silhouettes indicate internal consistency, not a mature or widely supported line of researce

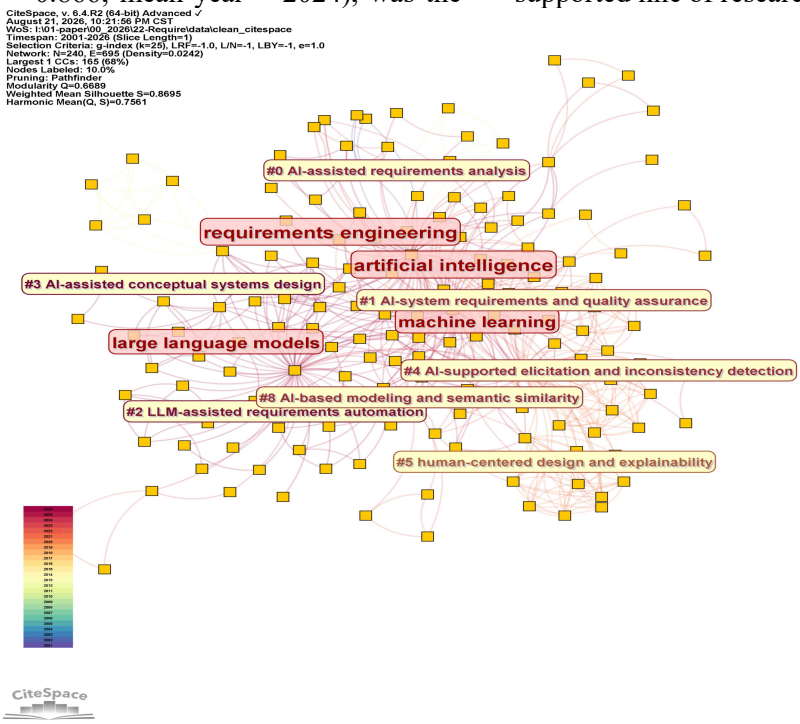

Figure 3. Keyword Co-Occurrence Network and Manually Calibrated Thematic Clusters

3.3 Thematic Evolution and Keyword Bursts

Requirements engineering extended across the longest period in the timeline (Figure 4). Terms related to ML and NLP occurred mainly in the middle and later time slices, while LLM-related

studies appeared at the most recent end of the map. Among the seven clusters, Cluster #2 had the latest mean year (2024). Its temporal position reflects the recent concentration of studies on LLM-assisted user-story assessment, specification, and development automation.

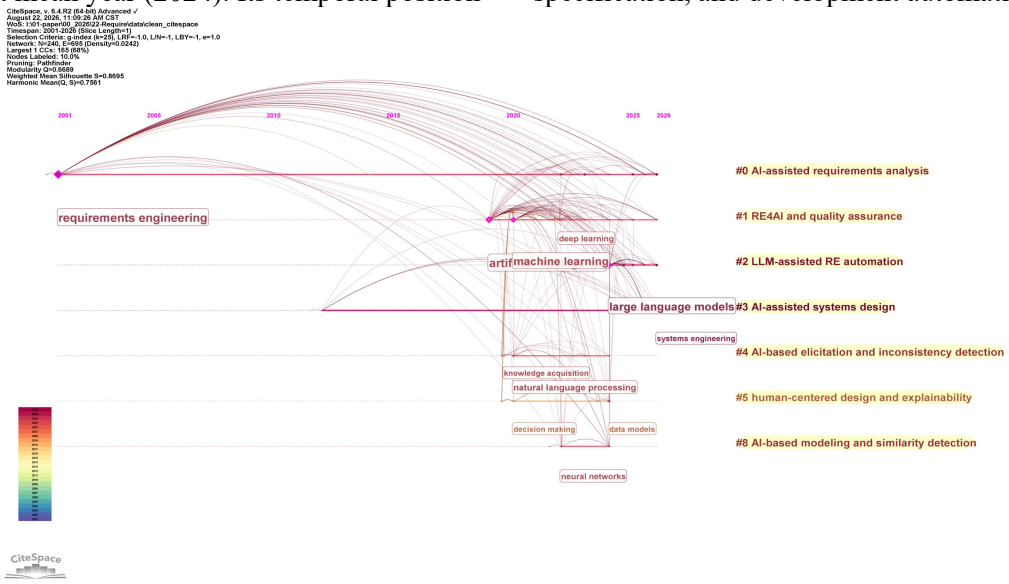

Figure 4. Timeline of the Seven Principal Keyword Clusters

Using the predefined minimum duration of two years, the burst analysis returned no keyword. A sensitivity run with the duration set to one year identified large language models with a burst strength of 4.9736, beginning in 2025 and continuing to the July 2026 endpoint. This result was not present under the primary setting, and the observation for 2026 covers only part of the year. It is therefore treated as a short-term signal rather than evidence of a sustained trend.

3.4 Reference Co-Citation Structure and Knowledge Bases

Reference co-citation analysis produced a network of 911 nodes and 952 links, with a density of 0.0023. Only 146 nodes belonged to the LCC, representing 16% of the full network. Changing the g-index parameter to k = 15 or k = 50 reduced LCC coverage to 14% and 11%, respectively; k = 25 was therefore used for the reported network. At this setting, modularity Q was 0.9267 and the weighted mean silhouette was 0.9916. The high values indicate clear separation and internal consistency within the connected portion of the network. They do not remove the broader fragmentation indicated by the 16% LCC coverage (Figure 5).

The analysis retained four co-citation clusters. The largest, Cluster #3, was labeled AI-enabled requirements analysis and automation (size = 49; silhouette = 0.983; mean year = 2019). It was also the most heterogeneous cluster, bringing together cited work on NLP for RE, requirements classification, LLMs in software engineering, and RE4AI mapping. Zhao et al. [1] was among the references with the joint-highest frequency and supplied a broad account of NLP-supported RE activities.

Cluster #4 was labeled ML-supported requirements classification and design guidance (size = 48; silhouette = 1.000; mean year = 2010). Supervised classification of functional and non-functional requirements formed an important part of its cited base [14]. Cluster #8, requirements-to-model transformation and ambiguity detection (size = 31; silhouette = 0.990; mean year = 2014), linked early natural-language analysis with model-driven transformation. The systematic language-analysis component was represented by Ambriola and Gervasi [15]. Because only two associated citing records supported each of these clusters, their labels characterize the local citing sets and should not be read as stable categories for the field as a whole. Cluster #12, requirements engineering for ML systems and responsible AI (size = 18; silhouette = 0.997; mean year = 2004), included surveys of engineering AI-based and ML-enabled systems [6] together with work on deployment, human, and ethical requirements. Its smaller size and mixed citing set warrant cautious interpretation. Table 2 summarizes the calibrated keyword and reference co-citation clusters and their representative sources.

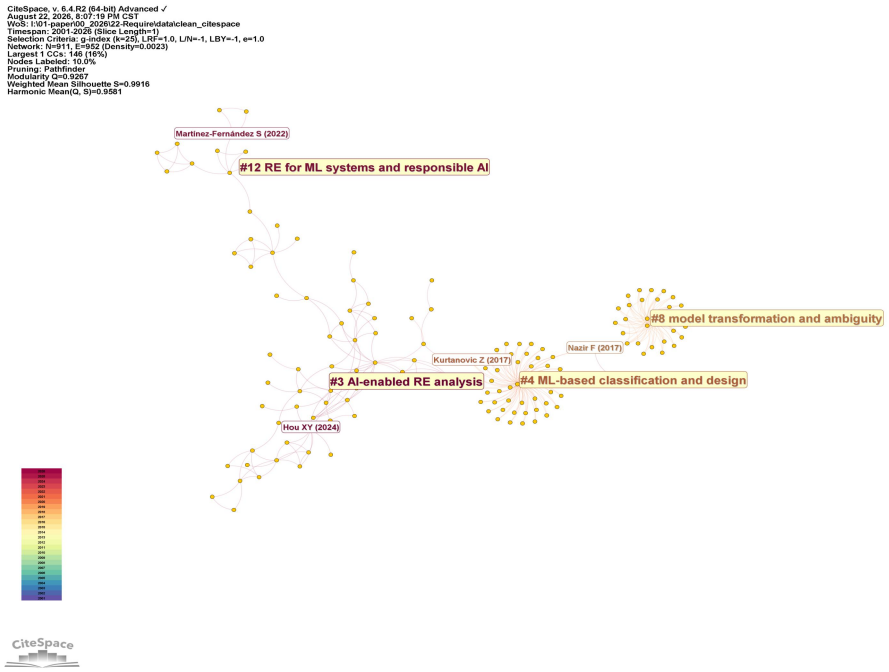

Figure 5. Reference Co-Citation Network and Manually Calibrated Knowledge-Base Clusters

Table 2. Major Keyword and Reference Co-Citation Clusters

Net.	ID	n	S	Year	Calibrated theme	Stream	Ref.
KW	#0	41	0.849	2020	AI-assisted RE analysis	Both	[10]
KW	#1	38	0.773	2022	RE4AI and quality assurance	RE4AI-led	[5]
KW	#2	34	0.866	2024	LLM-assisted RE automation	AI4RE	[2]
KW	#3	18	0.965	2021	AI-assisted systems design	Both	[3]
KW	#4	14	0.957	2020	Elicitation and inconsistency detection	AI4RE-led	[11,12]
KW	#5	13	0.960	2020	Human-centered explainability	Both	[7]
KW	#8	7	0.940	2022	Modeling and similarity detection	AI4RE	[13]
CC	#3	49	0.983	2019	AI-enabled RE analysis and automation	AI4RE-led	[1]
CC	#4	48	1.000	2010	ML classification and design guidance	AI4RE	[14]
CC	#8	31	0.990	2014	Requirements-to-model transformation	AI4RE	[15]
CC	#12	18	0.997	2004	RE for ML systems and responsible AI	RE4AI-led	[6]

Notes. KW = keyword; CC = reference co-citation. Labels were calibrated from representative records. KW #5 and #8 and CC #4 and #8 each had only two associated records. The co-citation-network LCC covered 16% of nodes.

3.5 Country and Regional Collaboration

The supplementary country/region network contained 50 nodes and 75 links (density = 0.0612), and its LCC covered 42 countries or regions (84%). Germany had the highest publication frequency (18), followed by the United States (12), Spain and South Korea (11 each), and Saudi Arabia and Sweden (9 each). Spain, Canada, Pakistan, England, and Italy had the highest unweighted degree (7 each). These counts describe participation and collaboration breadth, not national research quality or impact. Because this network added contextual rather than thematic evidence, it was retained as a supplementary result and not included among the five principal figures.

4. Discussion

4.1 From Task-Specific AI4RE to Generative Workflows

The maps reveal both continuity and a recent shift in AI4RE. NLP and supervised ML remain central to requirements classification, defect detection, stakeholder-text mining, traceability, and model extraction. These tasks have a longer evaluation history than LLM-based generation. Supervised classification and systematic language analysis have been studied across several datasets, and zero-shot learning directly addresses the shortage of labeled RE data [10, 13]. Online-forum mining similarly broadens the sources from which requirements knowledge can

be gathered [11]. Even so, results remain sensitive to datasets, requirement categories, and domains, which supports a moderate rather than established maturity rating.

The LLM cluster widens the scope of automation from bounded analytic tasks to generative, multi-stage workflows. Studies in this cluster use LLMs to evaluate user stories, prepare specification drafts, decompose functions, and generate downstream models or code [2, 3]. Their concentration in the latest time slices is also reflected in the one-year burst for large language models. Moving from classification to multi-stage generation changes the nature of the risk: an omission or hallucination at one stage may carry forward as an inconsistent transformation or a broken trace link. The agentic design study illustrates this limitation, because greater design detail did not fully resolve requirements coverage or code-fidelity problems. On the evidence currently available, LLMs are better treated as supervised aids for drafting and transformation than as substitutes for analyst judgment. Evaluation of these workflows also needs to cover provenance, deterministic verification, and explicit points for human review.

4.2 RE4AI as a Complementary Research Stream

RE4AI concerns the requirements of AI-enabled systems rather than simply the use of an AI tool during RE. Functional accuracy is only one part of this requirements problem. Data quality, explanations, stakeholder responsibilities, safety, fairness, trust, and post-deployment monitoring must also be considered. In the mapped corpus, these issues occur in the RE4AI and quality-assurance keyword cluster, the human-centered explainability cluster, and the responsible-AI co-citation cluster. This pattern is consistent with earlier mapping evidence that conventional RE approaches are not yet well adapted to AI systems [5]. It also accords with software-engineering surveys that report persistent difficulties during deployment and maintenance [6].

Some human-centered frameworks translate these broad concerns into requirements that can be discussed before implementation. ExplanaSC, for example, identifies the information users require when automated smart-contract decisions need to be explained [7]. Evidence for such frameworks, however, is often limited to a case application accompanied by expert feedback. Questions remain about how explanation, fairness, and trust requirements can be measured in a consistent manner, how conflicts among them should be resolved, and how they can be maintained after data or models change. Given these limits, the evidence synthesized here is classified as emerging to moderate, not established.

The overlap between AI4RE and RE4AI creates a further evaluation problem. If an AI method helps elicit or validate requirements for an AI-enabled product, requirements also apply to that supporting method, including traceability, reproducibility, security, and human control. Research located at this intersection needs to state clearly whether its evaluation concerns the RE task, the AI-enabled product, or the combined human-AI workflow.

Table 3. AI Approaches, RE Focus, Value, Limitations, and Evidence Maturity

Approach	Main RE focus	Stream	Reported value	Main limitation	Maturity
Supervised/transfer/zero-shot models	Classification	AI4RE	Scalable; fewer labels	Domain dependence; weak transfer	Moderate
Stakeholder-text mining	Elicitation	AI4RE	Broader evidence	Noise, bias, context loss	Moderate
Formal logic with LLMs	Contradiction checks	AI4RE	Explicit reasoning	Incomplete checks; controlled language	Emerging-moderate
Generative/agentic LLMs	Drafting and transformation	AI4RE	Rapid multi-stage support	Hallucination; weak traceability	Emerging
Similarity and NLP-to-model methods	Modeling and traceability	AI4RE	Links among artifacts	Template distortion; limited transfer	Moderate
Explainability frameworks	Information and explanation	RE4AI	Explicit user and explanation needs	Small cases; difficult trade-offs	Emerging
RE for ML/responsible AI	Data, safety, fairness, governance	RE4AI	Lifecycle coverage	Drift; contested values; accountability	Emerging-moderate

Note. Emerging denotes framework, proof-of-concept, or small-case evidence; moderate denotes a broader evidence base with limited external validation. No approach met the criterion for established evidence.

4.3 Research Gaps and Practical Implications

External validity and traceability are two recurring weaknesses in the mapped literature. Many approaches have been evaluated with a public dataset, generated requirements, a single case, or data from one organization. Evidence of transfer will require studies across projects and domains, with outcomes such as analyst effort, escaped defects, and downstream rework reported alongside model accuracy. Traceability presents a related problem. When an approach generates user stories, models, tests, or code, the resulting artifacts need links back to the source evidence, prompts, model versions, and human decisions that produced them.

The available evaluations also cover only a limited part of the system lifecycle. Changes in data, retraining, model replacement, and environmental drift can alter AI behavior after the initial requirements have been defined. RE4AI studies therefore need to connect those requirements with runtime monitoring, specified change triggers, and responsibility for corrective action. Human oversight requires similar design attention. An evaluation of oversight should examine automation bias, reviewer calibration, stakeholder participation, the usefulness of explanations, and who remains accountable when an AI-generated artifact is accepted.

For current practice, the findings support adoption on a selective basis. Classification, search, elicitation support, and consistency checking are most defensible when the task boundaries and validation criteria are explicit. Generative tools can shorten drafting or transformation work, but their use depends on grounding in source material, defined review checkpoints, and trace links that can be inspected. Independent assessment of research results also depends on reporting the datasets, model and prompt versions, baselines, observed failures, and human-review procedures.

Table 3 uses maturity ratings to keep research visibility separate from practical readiness. Cluster size or recency records the attention given to a topic; it does not demonstrate effectiveness across settings. None of the technique groups reviewed here had been evaluated in enough industrial contexts to justify an “established” rating. The distinction matters especially for LLM-based work. Although LLMs are highly visible in the recent maps, their evaluation base is newer and more varied than that available for conventional NLP and supervised ML.

4.4 Limitations

The review has five main limitations. First, it used one database and a deliberately broad query. Screening removed 900 retracted, concern-marked, or off-topic records, but studies using different terminology may have been missed. Second, the 36 borderline records were retained through bibliographic-field checks rather than dual-reviewer full-text screening; this is a bibliometric review, not a PRISMA-level systematic review. Third, the corpus was relatively small, and network density and cluster composition depend on CiteSpace settings. Parameter sensitivity was examined for the reference network, but not for every possible configuration.

Fourth, the co-citation network was highly fragmented: its LCC covered only 16% of nodes. High modularity and silhouette values therefore describe the connected clusters, not comprehensive coverage of the knowledge base. Several retained clusters also had only two associated corpus or citing records. Fifth, the 2026 data ended on 31 July. The LLM burst appeared only when the minimum duration was reduced to one year and should be treated as a sensitivity result rather than a forecast. Citation counts, indexing, and cluster labels may change as the literature develops.

5. Conclusion

Using 121 articles from the Web of Science Core Collection, this review examined the bibliometric structure of research connecting AI and RE. Keyword analysis resolved seven themes: AI-assisted analysis, RE4AI and quality assurance, LLM-assisted automation, systems design, elicitation and inconsistency detection, human-centered explainability, and AI-based modeling. The co-citation results linked these themes to knowledge bases in NLP-supported automation, ML classification, requirements-to-model transformation, and the engineering of responsible ML systems. At the same time, the low coverage of the largest connected component shows that the cited literature remains fragmented.

The results place AI4RE and RE4AI on related but different trajectories. AI4RE has moved from task-specific NLP and ML applications toward generative and agentic support at several

stages of RE. RE4AI broadens the requirements problem to include data, model behavior, safety, fairness, explainability, governance, and changes across the lifecycle. Where the two streams intersect, traceability, accountability, and human judgment need to be considered during design rather than after deployment. The studies reviewed here support selective use under human supervision, but they provide limited evidence for generalization across industrial settings or for performance over time.

Further work needs to test methods with cross-domain benchmarks, record model and prompt configurations, preserve links between generated artifacts and their source requirements, and evaluate governed human-AI workflows throughout the system lifecycle. Future bibliometric studies would also benefit from wider database coverage, independent full-text screening, systematic content coding, and sensitivity checks for additional parameter settings. Accordingly, the contribution of this study is a structural account of the screened corpus; it should not be read as a ranking of individual methods or as an exhaustive catalogue of AI-RE research.

Acknowledgments

This work was supported by the Fujian Province Young and Middle-aged Teachers' Education Research Project (Grant No. JAT220280) and the Fujian Province Natural Science Foundation (Grant No. 2025J01356).

References

[1] Zhao L, Alhoshan W, Ferrari A, Letsholo KJ, Ajagbe MA, Chioasca EV, Batista-Navarro RT. Natural Language Processing for Requirements Engineering: A Systematic Mapping Study. ACM Computing Surveys, 2021, 54(3): Article 55, 1-41.

[2] Sharma A, Tripathi AK. Evaluating user story quality with LLMs: a comparative study. Journal of Intelligent Information Systems, 2025, 63(4): 1423-1451.

[3] Massoudi S, Fuge M. Agentic Large Language Models for Conceptual Systems Engineering and Design. Journal of Mechanical Design, 2026, 148(5): 051405.

[4] Cheng H, Husen JH, Lu Y, Racharak T, Yoshioka N, Ubayashi N, Washizaki H. Generative AI for Requirements Engineering: A Systematic Literature Review. Software: Practice and Experience, 2026, 56(2): 141-170.

[5] Ahmad K, Abdelrazek M, Arora C, Bano M, Grundy J. Requirements engineering for artificial intelligence systems: A systematic mapping study. Information and Software Technology, 2023, 158: 107176.

[6] Martínez-Fernández S, Bogner J, Franch X, Oriol M, Siebert J, Trendowicz A, Vollmer AM, Wagner S. Software Engineering for AI-Based Systems: A Survey. ACM Transactions on Software Engineering and Methodology, 2022, 31(2): Article 37, 1-59.

[7] Al Ghanmi H, Bahsoon R. ExplanaSC: A Framework for Determining Information Requirements for Explainable Blockchain Smart Contracts. IEEE Transactions on Software Engineering, 2024, 50(8): 1984-2004.

[8] Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 2006, 57(3): 359-377.

[9] Chen C. Science mapping: A systematic review of the literature. Journal of Data and Information Science, 2017, 2(2): 1-40.

[10] Alhoshan W, Ferrari A, Zhao L. Zero-shot learning for requirements classification: An exploratory study. Information and Software Technology, 2023, 159: 107202.

[11] Khan JA, Liu L, Wen L. Requirements knowledge acquisition from online user forums. IET Software, 2020, 14(3): 242-253.

[12] Gärtner AE, Göhlich D. Automated requirement contradiction detection through formal logic and LLMs. Automated Software Engineering, 2024, 31(2): 49.

[13] Saini R, Mussbacher G, Guo JLC, Kienzle J. Automated, interactive, and traceable domain modelling empowered by artificial intelligence. Software and Systems Modeling, 2022, 21(3): 1015-1045.

[14] Kurtanović Z, Maalej W. Automatically Classifying Functional and Non-functional Requirements Using Supervised Machine Learning. 2017 IEEE 25th International Requirements Engineering Conference, 2017: 490-495.

[15] Ambriola V, Gervasi V. On the Systematic Analysis of Natural Language Requirements with CIRCE. Automated Software Engineering, 2006, 13(1): 107-167.

